Natural Language Descriptions of Human Behavior from Video Sequences

Carles Fernández Tena[1], Pau Baiget[1], Xavier Roca[1], and Jordi Gonzàlez[2]

[1] Computer Vision Centre, Edifici O. Campus UAB, 08193, Bellaterra, Spain
[2] Institut de Robòtica i Informàtica Ind. UPC, 08028, Barcelona, Spain
{perno,pbaiget,xroca,poal}@cvc.uab.es

Abstract. This contribution addresses the generation of textual descriptions in several natural languages for evaluation of human behavior in video sequences. The problem is tackled by converting geometrical information extracted from videos of the scenario into predicates in fuzzy logic formalism, which facilitates the internal representations of the conceptual data and allows the temporal analysis of situations in a deterministic fashion, by means of Situation Graph Trees (SGTs). The results of the analysis are stored in structures proposed by the Discourse Representation Theory (DRT), which facilitate a subsequent generation of natural language text. This set of tools has been proved to be perfectly suitable for the specified purpose.

1 Introduction

The introduction of Natural Language (NL) interfaces into vision systems is becoming popular, especially for surveillance systems. In these surveillance systems, human behavior is represented by scenarios, i.e. predefined sequences of events. The scenario is evaluated and automatically translated into text by analyzing the contents of the images over time, and deciding on the most suitable predefined event that applies in each case. Such a process is referred to as Human Sequence Evaluation (HSE) in [3]. HSE takes advantage of cognitive capabilities for the semantic understanding of human behaviors observed in image sequences.

This automatic analysis and description of temporal events was already tackled by Marburger et al. [7], who proposed a NL dialogue in German to retrieve information about traffic scenes. More recent methods for describing human activities from video images have been reported by Kojima et al. [6], and automatic visual surveillance systems for traffic applications have been studied by Nagel [8] and Buxton and Gong [2], among others. These approaches present one or more specific issues such as textual generation in a single language, surveillance for vehicular traffic applications only, restrictions for uncertain data, or very rigid environments, for example.

We aim to build a system which addresses the aforementioned drawbacks by following the proposals of HSE, in order to generate NL descriptions of human behavior appearing in controlled scenarios, for several selectable languages. Such

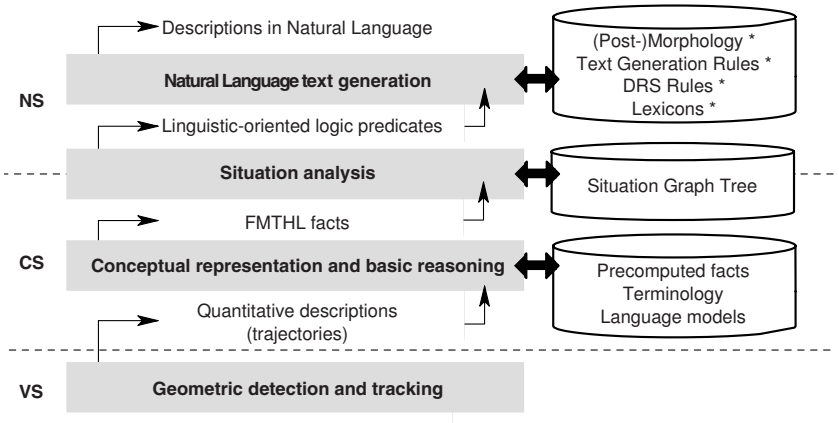
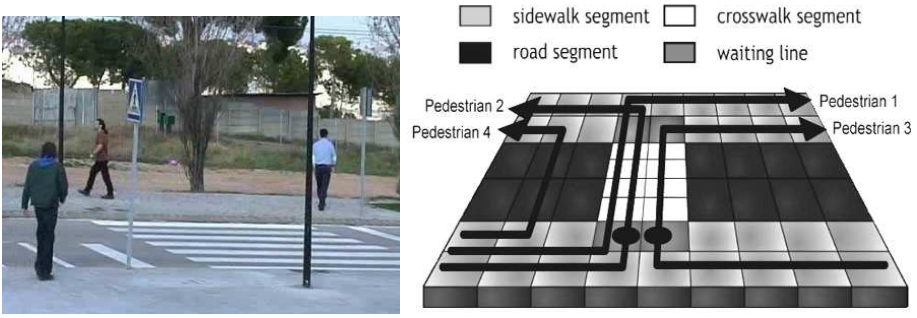


Fig. 1. General schema of the stages and interfaces related to the current text generation system. The left acronyms represent different sub-systems, the boxes describe the main processes that produce changes in data representations, and the right components specify some of the external tools required by the processes. An asterisk remarks that a resource is language-dependent.

a system builds upon three disciplines, namely computer vision, knowledge representation, and computational linguistics. Thus, the overall architecture consists of three subsystems, see Fig. 1; a Vision Subsystem (VS), which provides the geometric information extracted from a video sequence by means of detection and tracking processes, a Conceptual Subsystem (CS), which infers the behavior of agents from the conceptual primitives based on the geometric information extracted by the VS, and a Natural Language Subsystem (NS), which in principle comprises the NL text generation, but also becomes a good stage for providing a complete interface of communication with a final user [8]. Due to space limitations, the extraction of visual information is not treated here. Details can be found, for example, in [10]. We proceed on the basis that structural information consisting of geometrical values are available over time.

The obtention of knowledge derived from visual acquisition implies a necessary process of abstraction. In order to understand the quantitative results from vision, it becomes fundamental to reduce the information to a small number of basic statements, capable of detecting and relating facts by means of qualitative derivations from what has been 'seen'. The conversion of observed geometrical values over time into predicates in a fuzzy logic formalism allows to reach an intermediate state, the conceptual representation layer, which facilitates schematic representations of the scenarios [1] and, in addition, enables characterizations of uncertain and time-dependent data extracted from image sequences. Next, a classification can be performed by integrating these resulting facts into preconceived patterns of situations. Such an inference system produces not only an interpretation for the behavior of an agent, but also reasons for its possible reactions and predictions for its future actions [4].



(a) (b)

Fig. 2. Original pedestrian crosswalk scene (a) and groundplane schematic map of the main regions considered in this scene (b). Pedestrian trajectories have been included in the scheme. Black circles represent a stop on the waiting line.

Discourse Representation Theory seems to be of particular interest for the conversion from conceptual to linguistic knowledge, since it discusses algorithms for the translation of coherent NL textual descriptions into computer-internal representations by means of logical predicates [5]. The reverse step is also possible, so that the results of the conceptual analysis are stored into semantic containers, the so-called Discourse Representation Structures (DRS), which facilitate the construction of syntactical structures containing some given semantic information. A final surface realization stage over these preliminary sentences embeds the morphological and orthographical features needed for obtaining final NL textual descriptions.

Next chapter describes the chosen scenario, and explains how the evaluation of human behaviors is achieved from spatiotemporal data and prior knowledge about the scene. Section 3 details the mechanisms which convert high-level predicates obtained from situational analysis into NL textual descriptions. Some experimental results for Catalan, English, and Spanish are shown in Section 4. Finally, Section 5 concludes the paper and suggests future lines of work.

2 Evaluation of Human Behaviors in Video Sequences

The chosen scenario for evaluation of basic human behaviors has been a crosswalk, see Fig. 2. On it, a certain number of pedestrians, each one with a different behavior, start from one of the sidewalks and cross the road to get to the other side. At first, the presence of traffic vehicles has been omitted.

2.1 The Conceptualization Step

The structural knowledge acquired by the VS needs to be abstracted and converted into logic knowledge in order to facilitate further manipulations and reasonings. To do so, trajectories and other types of estimated spatiotemporal

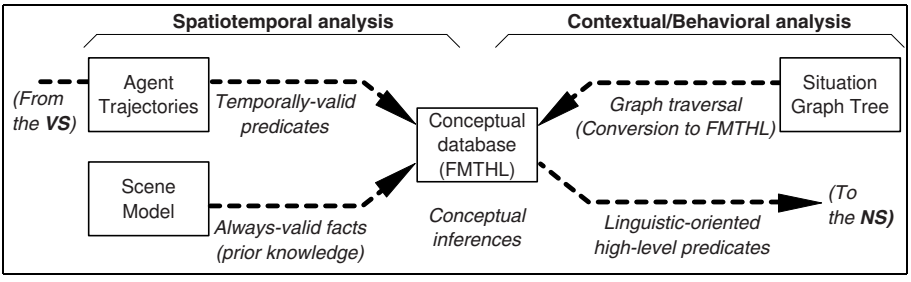


Fig. 3. Quantitative data (e.g. trajectories from the agents) is evaluated upon the predefined facts from the scene model, and thus converted into qualitative FMTHL knowledge. On the other hand, the behavioral model encoded into a SGT is traversed and converted into FMTHL conditions, too. Finally, the entire set of asserted spatiotemporal qualitative results is logically classified by the traversed SGT, and *linguistic-oriented predicates* are generated as a result.

information are associated to basic concepts identifying recognizable simple actions, which can be described by using elementary verb-phrases (e.g., 'approaching_to_location', 'turning', 'has_speed'). These *conceptual predicates* are not yet proper linguistic expressions, but system-internal representations resulting from classification and abstraction processes.

Fuzzy Metric Temporal 'Horn' Logic (FMTHL) has been conceived as a suitable mechanism for dealing with uncertain, time-dependent information [11]. This formalism allows to represent knowledge explicitly and hierarchically, not coded into conditional probabilities, and enables to manage data requiring both *temporal* and *fuzzy* properties [4]. In our case, observed trajectories from the agents are analyzed within a predefined scene model, see Fig. 3. As a result, geometrical, quantitative values are acquired, such as postures, velocity, or positions for the agents. After an abstraction process is carried out, the reasoning system is conferred a capability for representing uncertain qualitative descriptions inferred from the quantitative data. The logic productions evolve over time as the received data does, so this conceptual knowledge is also time-delimited, and thus the development of events can be comprehended and even anticipated.

The qualitative knowledge extracted from quantitative results is encoded using these fuzzy membership functions, so that the generated predicates are related to conceptual 'facts' for each time-step. For example, a collection of positions over time allows to derive fuzzy predicates such as 'has_speed(zero)', 'has_speed(small)', or 'has_speed(very_high)', depending on the displacements of the agent detected between consecutive points of time.

2.2 Agent Trajectories

Trajectory files are ordered collections of observed values over time for a certain agent, which are obtained as a result of the tracking processes for the agents [10]. From the evolution of the states of the agent, a certain *behavior*, i.e. a sequence

of situated actions, will be assumed. Four agent trajectories have been obtained, which consist of a set of FMTHL logical predicates of type `has_status`. These predicates comprise the required knowledge for the human behavior analysis in the following scheme or *status vector* for the agent at time `t`:

```
t !    has_status(Agent, X0, Y0, Theta, Vel).
```

As can be seen, the `has_status` predicates for the interpretation of human actions contain five fields so far, all of them being identifiers to entities and objects detected during the tracking process, or otherwise concrete geometrical values for spatiotemporal variables. The `Agent` field gives information about the name given to the agent. The rest of the fields give quantitative values to the geometrical variables needed: 2-D spatial position in the ground plane (`X0, Y0`), angle of direction (`Theta`), and instant velocity (`Vel`). The `Vel` field provides the necessary information for determining the action being performed by the agent (i.e. *standing*, *walking*, *running*).

2.3 Scene Modeling

The scenario in which pedestrians perform their actions has been included as an additional source of knowledge for the reasoning stage. The geometrical modeling of the location has been done first in a ground plane bidimensional approach, so a set of spatial descriptors are declared to distinguish the relevant topographic or interesting elements in the scene, see Fig. 2 (b). This source provides the spatial distribution taken into account for the given situation.

A second source of knowledge contains other logical statements that will confer semantic significance on the initial geometrical descriptors of the scene. The different regions can be enclosed into different categories (*sideway*, *road*, *crosswalk*) and can also be given different attributes (*walking zone*, *waiting line*, *exit*). This step is necessary for identifying significative regions, so the movements and interactions of the agents can be contextualized by means of valid identifiers. These geometrical considerations have been encoded using FMTHL predicates.

2.4 Situation Graph Tree

Situation Graph Trees (SGTs) are hierarchical structures used to model the knowledge required from human behavior analysis in a specific discourse domain [3]. A SGT has been designed for the crosswalk scene, see Fig. 4. The conceptual knowledge about a given actor for a given time step is contained in a so-called *situation scheme*, which constitutes the basic components of a SGT. The knowledge included in these components is organized in two fields: *state predicates* and *action predicates*.

- First, a set of logic conditions describes the requirements that need to be accomplished to instantiate that situation. The assertion of these *state predicates* is performed by evaluating the semantic predicates inferred from the agent status vectors obtained at the visual stage.

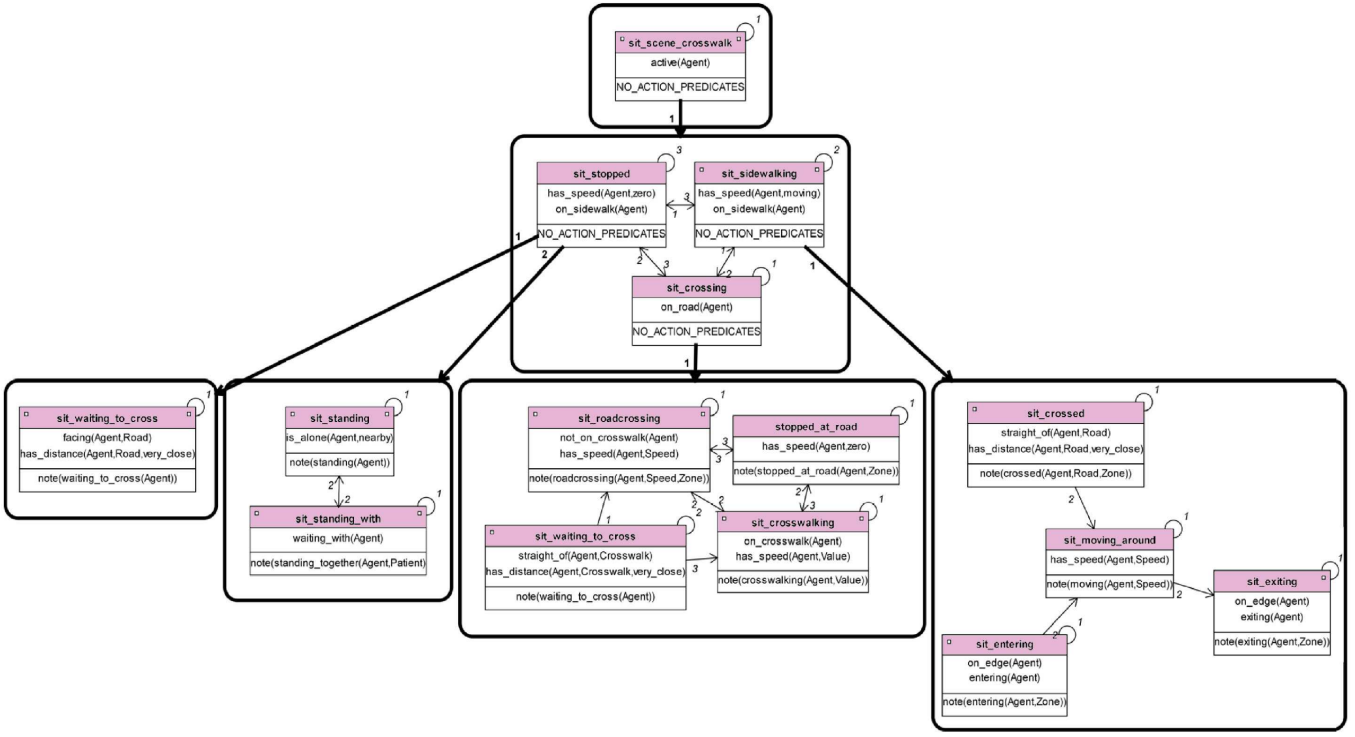


Fig. 4. Situation graph tree describing the behaviors of pedestrians on a crosswalk. Situation graphs are depicted as rounded rectangles, situation schemes are shown as normal rectangles. Bold arrows represent particularization edges, thin arrows stand for prediction edges, and $\frac{3}{4}$–circle arrows indicate self-predictions. Small rectangles to the left or to the right of the name of situation schemes mark that scheme as a start- or end-situation, respectively [1]. A SGT needs to focus on *behaviors* of the agents, while avoiding dependance to a particular scenario. The more this approach is achieved, the more flexible and scene-independent the SGT will be.

– After the conditions have been asserted, certain domain-specific *action predicates* are generated and forwarded for defined purposes. Only generation of NL text will be considered here, so *linguistic-oriented* logic predicates will be generated (*note* commands in Fig. 4).

A single SGT incorporates the complete knowledge about the behavior of agents in a discourse [1]. Every possible action to be detected has to be described in the SGT. Consequently, it is necessary to have accuracy to precisely identify the desired actions, but it is also important that it does not become excessively complex in order to avoid a high computational cost. On the other hand, the SGTs are transformed into logic programs of a FMTHL for automatic exploitation of these behavior schemes, as shown in Fig. 3.

Depending on the behavioral state, a new high-level predicate will be sent to the NS Subsystem, by means of a `note` method. The new predicates offer language-oriented structures, since their attribute scheme comprises fields related to ontological categories such as *Agent*, *Patient*, *Object* or *Event*. These predicates are the inputs for the NS Subsystem, which will be discussed next.

3 Linguistic Implementation

It is in the NS where the logical predicates are used to provide the representational formalism, making use of the practical applications of the Discourse Representation Theory. Inside the NS layer, there are several stages to cover:

1. Lexicalization
2. Discourse Representation
3. Surface Realization

Besides, the set of lemmata been used has to be extracted from a restricted corpus of the specific language. This corpus can be elaborated based upon the results of several psychophysical experiments on motion description, collected over a significative amount of native speakers of the target language. In our case, ten different people have independently contributed to the corpus with their own descriptions of the sample videos. Three different languages have been implemented for this scenario: Catalan, English, and Spanish.

3.1 Generation of textual descriptions

The overall process of generation of NL descriptions is based on the architecture proposed by Reiter & Dale [9], which includes three modules; a document planner, a microplanner, and a surface realizer (see Fig. 5). The VS provides the information to be communicated to the user; this task is considered to be part of the Document Planner. The CS decides how this information needs to be structured and gives coherency to the results. This module provides general reasoning about the domain and determines the content to be included in the sentences to be generated, which are tasks related to the Document Planner, too. Further tasks, such as microplanning and surface realization, are included into the NS. An example for the entire process of generation is shown in Fig. 6.

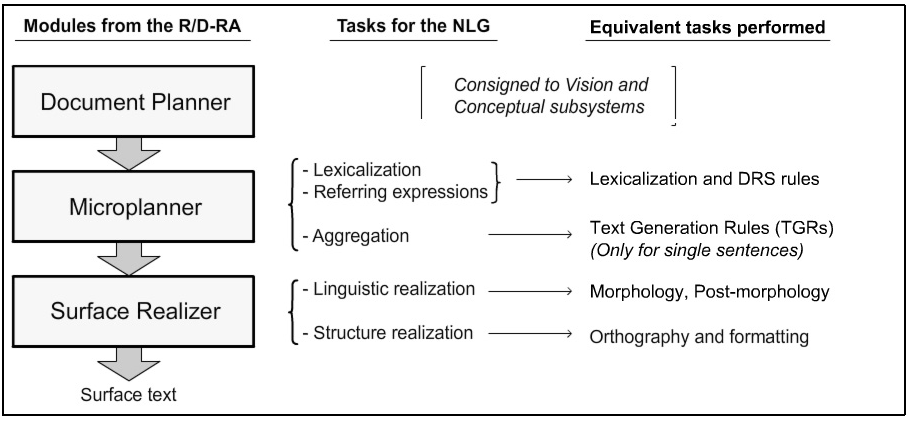


Fig. 5. Schema of Reiter/Dale Reference Architecture (R/D-RA) [9], including the tasks related to each module that are necessary for a Natural Language Generator

Lexicalization. It is necessary to convert the abstract predicates from the CS into linguistic entities for communication, such as agents, patients, objects, or events, for instance. The classification of linguistically-perceived reality into thematic roles (e.g. agent, patient, location) is commonly used in contemporary linguistic-related applications as a possibility for the representation of semantics, and justifies the use of computational linguistics for describing content extracted by vision processes. The lexicalization step can be seen as a *mapping process*, in which the semantic concepts identifying different entities and events from the domain are attached to linguistic terms referring those formal realities. This way, this step works as a real dictionary, providing the required lemmata that will be a basis for describing the results using natural language.

Representation of the Discourse. Nevertheless, bridging the semantic gap between conceptual and linguistic knowledge cannot be achieved only with a lexicalization step. Discourse Representation Structures (DRSs) are the actual mechanism that facilitates to overcome the intrinsic vagueness of NL terms, by embedding semantics inferred at the conceptual level into the proper syntactical forms. Lemmata are just units that will be used by these structures to establish the interrelations which will convey the proper meaning to the sentences.

DRSs are semantic containers which relate referenced conceptual information to linguistic constructions [5]. A DRS always consists of a so-called *universe* of referents and a set of conditions, which can express characteristics of these referents, relations between them, or even more complex conditions including other DRSs in their definition. These structures contain linguistic data from units that may be larger than single sentences, since one of the ubiquitous characteristics of the DRSs is their semantic cohesiveness for an entire discourse.

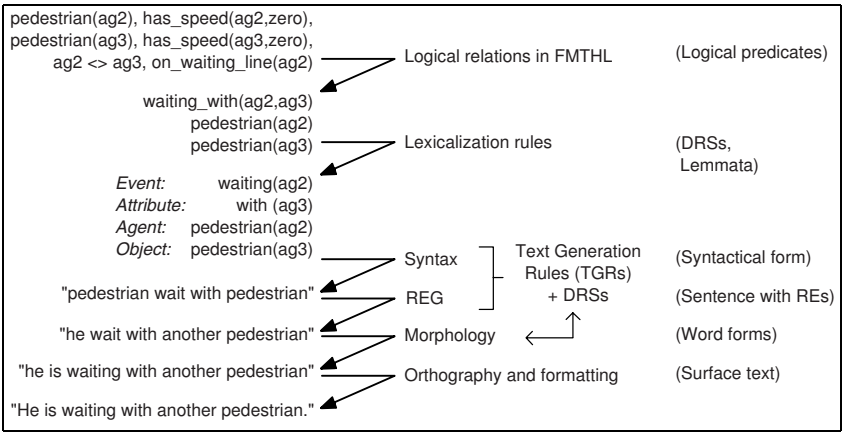


Fig. 6. Example for the generation of the sentence ''*He is waiting with another pedestrian*" from logical predicates. The center column contains the tasks being performed, and the right column indicates the output obtained after each task.

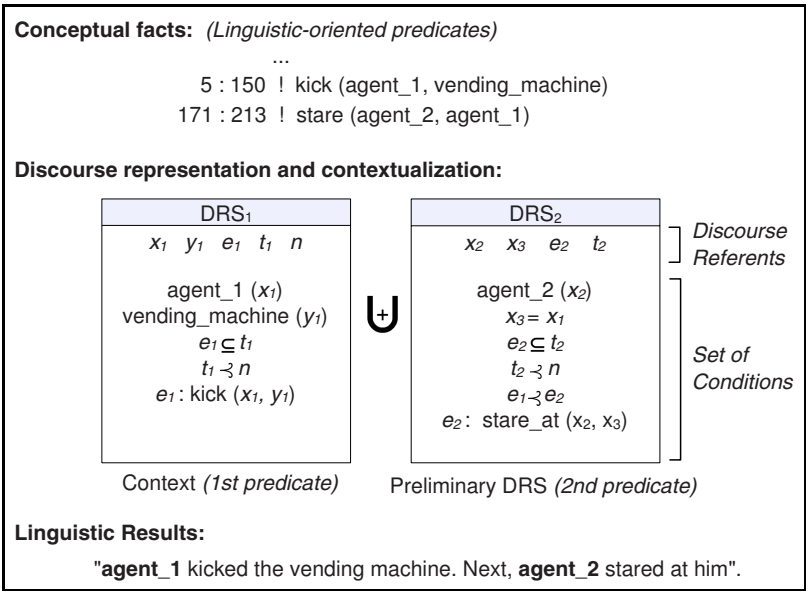


Fig. 7. A pattern DRS allows to convert a stream of FMTHL into a string of textual symbols. Here, two predicates are validated. The first one instantiates a DRS, which serves as context for the following asserted facts. Once the new predicate is validated, it instantiates another DRS which merges with that context, thus providing a new context for subsequent facts. The temporal order of the events is stated by relating them to time variables ($e_1 \subseteq t_1$), placing these variables in the past ($t_1 \prec n$), and marking precedence ($e_1 \prec e_2$).

When a contextual basis is explicitly provided, the maintenance of the meaning for a discourse, including its cross-references, relations and cohesion can be granted. Then, linguistic mechanisms such as anaphoric pronominalization for referring expressions can be successfully implemented, e.g. *'The pedestrian is running'* → *'He is running'*. In our case, since situational analysis is performed individually for every detected agent, we base on previously mentioned information about the focused agent to decide whether to use pronouns or full descriptions. An example which shows how the semantic representation and contextualization is undertaken by a DRS is illustrated in Fig. 7.

DRSs facilitate the subsequent tasks for sentence generation. The syntactical features of a sentence are provided by the so-called Text Generation Rules (TGRs), which establish the position for the elements of the discourse within a sentence for a particular language. Due to the specific goals considered for this system, several assumptions have been taken: we use simple sentences for effective communication

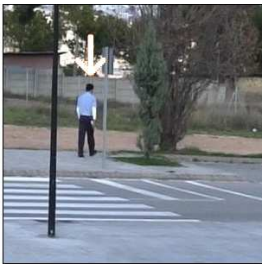
Surface Realization. The Surface Realization stage is accomplished in two steps. A first morphological process applies over each single word and partially disambiguates the individual abstraction of that word, by means of morphological attributions such as gender or number. These attributions can be propagated upon the semantic relations previously established by DRSs among the lemmata of a single piece of discourse. After that, a set of post-morphological rules has been conceived to enable interactions among predefined configurations of words, thus affecting the final surface form of the text. This additional step is indispensable for many languages, in which certain phenomena force the surface form to change, e.g. contractions (*'a'+'el'→'al'* in Spanish), or order variation (*'es'+'va'+'en'→'se'n va'* in Catalan).

4 Experimental Results

We address the problem in-depth for a particular domain, instead of finding a generically-applicable solution. For this reason, an ad-hoc solution has been chosen for the identification of a predefined set of behaviors in the described scenario. In such a framework, situations are specialized as long as spatiotemporal information can be classified by the given models. If a non-modeled situation occurs, the SGT cannot specialize a concrete interpretation, and instead of this it generates a more general description, e.g. pedestrians being detected in certain regions, or agents grouping or splitting. A group of native speakers provided linguistic interpretations for the set of behaviors, for each individual language considered.

Thus, the coverage of the generated descriptions is tightly related to the extent of situations modeled by the SGT. A vertical growing of this classifier, i.e. an increment of the particularization edges, increases the granularity of the descriptions, thus disambiguating or specializing the discourse. On the other hand, by enhancing the human motion models for the scene and the prediction edges, we





Pedestrian 3 (Catalan)

203 : *Lo vianant surt per la part inferior dreta.*
252 : *Va per la vorera inferior.*
401 : *S'espera per creuar.*
436 : *S'està esperant amb un altre vianant.*
506 : *Creua pel pas zebra.*
616 : *Va per la vorera superior.*
749 : *Se'n va per la part superior dreta.*

Pedestrian 3 (English)

203 : *The pedestrian shows up from the lower right side.*
252 : *He walks on the lower sidewalk.*
401 : *He waits to cross.*
436 : *He is waiting with another pedestrian.*
506 : *He enters the crosswalk.*
616 : *He walks on the upper sidewalk.*
749 : *He leaves by the upper right side.*



Pedestrian 4 (Spanish)

523 : *El peatón aparece por la parte inferior izquierda.*
572 : *Camina por la acera inferior.*
596 : *Cruza sin cuidado por la calzada.*
681 : *Camina por la acera superior.*
711 : *Se va por la parte superior izquierda.*

Pedestrian 4 (English)

523 : *The pedestrian shows up from the lower left side.*
572 : *He walks on the lower sidewalk.*
596 : *He crosses the road carelessly.*
681 : *He walks on the upper sidewalk.*
711 : *He leaves by the upper left side.*

Fig. 8. Some of the descriptions in NL which have been generated for the crosswalk scene. The results match perfectly with the purposed set of natural language sentences suggested by a group of native speakers of the given languages.

increase the number of possible situations and their temporal structure respectively. Finally, the quality of the textual corpora provided by native speakers, which link linguistic patterns to the conceived situations, determines the goodness of the discourse representation. A set of deterministic linguistic rules were designed so that results matched perfectly with the selection of the descriptions provided by native users.

Some results for the situation analysis of the crosswalk scene are shown in Fig. 8. Textual descriptions in Catalan, English, and Spanish have been selected for Agents 3 and 4, respectively. These descriptions include agents appearing or leaving the scene, interactions between pedestrians and locations within the scenario (crosswalk, sidewalks), and interpretations for some detected behaviors, such as waiting with other agents to cross, or crossing in a dangerous way (i.e. directly by the road and not caring for vehicular traffic). Only static cameras were used in this first step, so no expressions concerning the action of the cameras are generated. Next improvements should focus on the semantic content provided by the behavioral and inference subsystems, i.e. which situations must be considered for a certain domain and scenario, and in which way the reasonings for these situations have to be done. Further approaches will lead to more complex requirements regarding linguistic capabilities, which have been restricted so far.

5 Conclusions

A system that evaluates video sequences involving human agents by generating NL descriptions in multiple languages has been successfully developed in a first stage. A brief overview of the tasks performed may help to understand how the generated text contributes to the goal of human behavior evaluation. After the conceptualization of the spatiotemporal information is achieved, and basic inferences are done, SGTs are in charge of integrating the deduced semantic knowledge. Also, contextual and behavioral models are applied here, since SGTs can be seen as actual classifiers of content for situations in a definite domain. The generation of NL is built upon the high-level semantic predicates generated by a SGT. In some way, the generated descriptions are *interpretations* of this semantic knowledge accomplished by native speakers of a certain language. The group of native speakers choose the linguistic expressions they find more appropriate, in order to incorporate the situations from a SGT into a suitable discourse. Hence, the situations appearing in a video sequence from a given domain can be interpreted and described in multiple natural languages.

The current NS allows for a flexible and fast incorporation of languages into a facility for multilingual generation of textual descriptions in NL. The natural language formalism makes possible to generate fluid rich sentences to the user, allowing for detailed and refined expressions that are not possible by using other mechanisms. The interconnection of all the stages involved in the system has been proved as convenient for the whole process of evaluation, although several gaps still have to be solved. Further steps should include the extension of current behavioral models, the detection of groups and more complex interactions

among agents and/or vehicles, and the use of uncertainty for not only predicting behaviors, but also to enhance possible hypothesis of interpretation for the detected events within the scene.

Lastly, results from NL texts can be interpreted as semantic tags to provide content segmentation of the video sequences over time. We are currently studying the connection of a user interaction stage accepting input NL-based queries to a large database of video sequences, generic or specific. This will be the starting point for search engines capable of retrieving video sequences showing specific motion or factual contents. In addition to this, the segmentation of video sequences into time-intervals showing cohesive information can be applied for extracting a collection of few semantic shots from these sequences. This way, a compression of the relevant information – user-definable and freely configurable by declaring attentional factors – can be done by summarizing the entire videos with a list of behavior concepts. Thus, we aim to improve motion description patterns for video standards such as MPEG-7, thus allowing for high-level annotations related to the motion within the scene.

Acknowledgements

This work has been supported by EC grant IST-027110 for the HERMES project and by the Spanish MEC under projects TIC-2003-08865 and DPI-2004-5414. Jordi Gonzàlez also acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

1. Arens, M., Nagel, H.H.: Representation of Behavioral Knowledge for Planning and Plan–Recognition in a Cognitive Vision System. In: Jarke, M., Koehler, J., Lakemeyer, G. (eds.) KI 2002. LNCS (LNAI), vol. 2479, pp. 268–282. Springer, Heidelberg (2002)
2. Buxton, H., Gong, S.: Visual surveillance in a dynamic and uncertain world. AI-magazine 78(1), 431–459 (1995)
3. Gonzàlez, J.: Human Sequence Evaluation: The Key-Frame Approach. PhD thesis, Universitat Autònoma de Barcelona, Barcelona, Spain (2004)
4. Haag, M., Theilmann, W., Schäfer, K., Nagel, H.H.: Integration of Image Sequence Evaluation and Fuzzy Metric Temporal Logic Programming, pp. 301–312. Springer, London, UK (1997)
5. Kamp, H., Reyle, U.: From Discourse to Logic. Kluwer Academic Publishers, Dordrecht, Boston, London (1993)
6. Kojima, A., Tamura, T., Fukunaga, K.: Natural language description of human activities from video images based on concept hierarchy of actions. International Journal of Computer Vision 50(2), 171–184 (2002)
7. Marburger, H., Neumann, B., Novak, H.J.: Natural Language Dialogue about Moving Objects in an Automatically Analyzed Traffic Scene. In: Proc. IJCAI-81, Vancouver (1981)
8. Nagel, H.H.: Steps toward a Cognitive Vision System. AI-Magazine 25(2), 31–50 (2004)

9. Reiter, E., Dale, R.: Building Natural Language Generation Systems. Cambridge University Press, Cambridge/UK (2000)
10. Rowe, D., Rius, I., Gonzalez, J., Villanueva, J.J.: Improving Tracking by Handling Occlusions. In: Singh, S., Singh, M., Apte, C., Perner, P. (eds.) ICAPR 2005. LNCS, pp. 384–393. Springer, Heidelberg (2005)
11. Schäfer, K., Brzoska, C.: F-Limette Fuzzy Logic Programming Integrating Metric Temporal Extensions. Journal of Symbolic Computation 22(5-6), 725–727 (1996)